

# **Scientific and Technical Report**

Sponsored by  
Advanced Research Projects Agency/ITO  
and United States Patent and Trademark Office

Browsing, Discovery and Search in Large Distributed Databases  
of Complex and Scanned Documents

ARPA Order No. D468

Issued by EXC/AXS under Contract #F19628-95-C-0235

Date Submitted: January 15, 1997

Period of Report: October 1, 1996 to December 31, 1996

Submitted by: Professor W. Bruce Croft, Principal Investigator  
Computer Science Department  
University of Massachusetts, Amherst

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

Distribution Statement A: Approved for public release; distribution is unlimited.

19970121 117

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE 1/15/97	3. REPORT TYPE AND DATES COVERED Scientific/Tech 10/1/96 - 12/31/96		
4. TITLE AND SUBTITLE Browsing, Discovery, and Search in Large Distributed Databases of Complex and Scanned Documents			5. FUNDING NUMBERS F19628-95-C-0235 ARPA Order No. D468	
6. AUTHOR(S) W. Bruce Croft				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Massachusetts, Amherst Box 36010, OGCA, Munson Hall Amherst, MA 01003-6010			8. PERFORMING ORGANIZATION REPORT NUMBER TR528181197	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Mr. Harry Koch ESC/AXS Bldg 1704, Room 114 5 Eglin St. Hanscom AFB, MA 01731-2116			10. SPONSORING/MONITORING AGENCY REPORT NUMBER Ms. Monique Dillon Office of Naval Research Boston Regional Office 495 Summer St., Room 103 Boston, MA 02210-2109	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Distribution Statement A: Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words)  This project aims to integrate powerful, new techniques for interactive browsing, discovery, and retrieval in very large, distributed databases of complex and scanned documents. Emphasis is placed on going beyond full-text retrieval techniques developed in the DARPA TIPSTER program to support different types of access and non-textual content. These techniques should be particularly relevant to the patent domain where it is important to find relationships between documents and where the patent or trademark may be based on a visual design. The specific tasks identified involve studying representation techniques for long documents with complex structure, browsing and discovery techniques for large text databases, image retrieval and scanned document retrieval techniques, and architectures for large, distributed databases.				
14. SUBJECT TERMS Browsing      Query Processing      Indexing Image Retrieval      Scanned Document Retrieval      Bayesian Network Text Retrieval      Probabilistic Retrieval Model      Large Distributed Databases			15. NUMBER OF PAGES 9	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited	

## Table of Contents

Task 1: Representation techniques for Complex Documents.....	1
Task 2: Browsing and Discovery Techniques for Document Collections.....	2
Task 3: Scanned Document Indexing and Retrieval.....	3
Task 4: Distributed Retrieval Architecture.....	4

# **Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents**

## **Technical and Scientific Report**

### **Task 1: Representation Techniques for Complex Documents**

#### **Task Objectives**

In this task, the goal is to extend the word-based representations that are common in retrieval systems in order to support summarization, browsing, and more effective retrieval. Specifically, we will be studying phrase-based representations and relationships between phrases in individual and groups of documents as the basis for our approach. Document structure will be used as part of the information that is used to "tag" the phrasal representation.

#### **Technical Problems**

The technical problems have to do with defining a "phrase", developing techniques for rapidly extracting them from text, comparing phrase contexts to identify significant relationships, producing summaries from these representations, and extending the underlying retrieval model to be able to make effective use of phrasal representations in both query-based retrieval and relevance feedback.

#### **General Methodology**

The general methodology for this task is to demonstrate effectiveness through user-based and collection-based experiments. Extensive use will be made of the TIPSTER document collection, which consists of a large number of text documents from a variety of sources, queries, and user relevance judgments for each query. This collection will be used for the experiments involving new probabilistic retrieval models and relevance feedback. Summarization techniques will be compared to sentence-based approaches and user-based evaluations of these summaries will be done. As more work is done on summarization in the TIPSTER program, we will make use of any new evaluation measures developed there.

#### **Technical Results**

We have continued to refine the techniques for identifying phrases for indexing in large corpora. Based on our analysis of results from the TREC corpus, patterns that occur within fixed window sizes are not superior to patterns made up of consecutive words. In addition, rules for pruning the patterns and special classes of patterns (such as quantities) have been identified. An initial approach to incorporating large phrase lists in the indexing process has also been developed. We have also begun to look at phrase clustering as a technique for summarizing document collections and large documents. Global measures of significance that go beyond simple frequency counts are also being considered. Experiments on selecting the core concepts in a query have produced mixed results. In discussions with PTO, we have found that patent classification is an area of significant interest and we have begun to develop classification research tasks.

## Important Findings and Conclusions

It is too early in the project to claim any conclusive results. The TREC results showed no significant improvement obtained from the technique we were using to identify core terms, although this was not the case in earlier experiments. We are currently working on improving these techniques.

## Significant Hardware Development

Purchases were made of disk to store large collections of patent data.

## Special Comments

We continue to work with the PTO, San Diego Supercomputer consortium and DARPA to obtain access to some of the patent collections and establish fast network links in order to be able to use the very large archives of scanned patents.

## Implication for Further Research

The key to more experiments is to obtain patent data. We plan to refine the initial research goals based on feedback from PTO and present these in the next status report.

## **Task 2: Browsing and Discovery Techniques for Document Collections**

### Task Objectives

The goals of this task are to develop techniques for summarizing collections of documents, and discovering connections between important ideas and documents in distributed collections. These techniques will be designed to support interactive browsing in environments like the PTO.

### Technical Problems

The technical problems involve producing an effective summary of a group of documents, such as a retrieved set or an entire database. Both document and phrase clusters could be used as part of this process. In order to support discovery, connections must be made between documents and groups of phrases that use a variety of evidence in addition to direct co-occurrence.

### General Methodology

The techniques will be evaluated with user-based and collection-based experiments. The relevance judgments from the TIPSTER collection will be used to evaluate clusters of documents. Phrase clusters will be evaluated by their impact on retrieval effectiveness and through user experiments that will measure performance on specific tasks. Part of the effort in this task (and the previous one) will involve developing a PTO test collection, which means that sample queries will need to be gathered from patent examiners and they will need to evaluate demonstrations of tools as they are developed.

### Technical Results

We have continued to develop a 3-D graphics interface designed for manipulating document and concept relationships to identify strong groupings and relationships. This work is

being evaluated by comparing the visualization to known relevant groups based on TREC data. We have also started work on phrase clusters as summaries of retrieved sets and document collections.

#### Important Findings and Conclusions

Our experiments with the visualization tool have not yet produced significant results, but new approaches have shown some promise.

#### Significant Hardware Development

None

#### Special Comments

#### Implication for Further Research

We will continue to work on visualization and summarization techniques that are relevant to tasks in the PTO. Elements such as the PTO classification structure will play more of a role in future developments.

### **Task 3: Scanned Document Indexing and Retrieval**

#### Task Objectives

The goals of this task are to develop techniques for detecting text, trademarks, logos, and images in scanned documents, clean up backgrounds of these detected objects, and support retrieval of images (such as designs in design patents), trademarks, and text from OCR.

#### Technical Problems

Current zoning techniques available with commercial OCR devices do not accurately locate text or trademarks within other images. We are developing techniques based on gaussian derivative filters to both detect and clean up (remove noisy backgrounds) these classes of objects in scanned documents. We are developing "appearance-based" retrieval of images as well as more straightforward features such as color and texture. Filter based and frequency domain based techniques offer some potential in this area, but significant work needs to be done on making this approach efficient enough to deal with thousands of images.

#### General Methodology

The evaluation of these techniques will be done in a similar way to text by developing test collections of images and scanned documents. Specifically, we are working to obtain large collections of trademarks and design patents, as well as typical queries.

#### Technical Results

We have obtained initial results from indexing images and performing retrieval tests using a set of image-based queries. These results have shown that retrieval speed has improved by more than a factor of 20 while retrieval effectiveness has been maintained. We have also been developing test databases for text detection and initial evaluations of these techniques

look promising. We have begun to study retrieval of design patents, but this is still at an early stage.

#### Important Findings and Conclusions

Recent results indicate that our techniques can provide effective and efficient retrieval of some types of images. We must now apply these techniques to design patents and trademarks, and continue to work on indexing techniques that will support massive scaleup of image database size.

#### Significant Hardware Development

Disk acquisition.

#### Special Comments

Gaining access to PTO design patents and trademarks has been a priority.

#### Implication for Further Research

We plan to have more meetings with patent examiners who deal with design patents and trademarks. As soon as enough test data is obtained, we will begin work on customizing the techniques for the types of drawings found in design patents.

### **Task 4: Distributed Retrieval Architecture**

#### Task Objectives

The goals of this task are to scale up our current methods of automatically selecting collections and merging results, and to investigate architectures that can support efficient retrieval, browsing and relevance feedback in distributed environments with terabytes of information.

#### Technical Problems

The current INQUERY text retrieval system uses a client server architecture to support simultaneous retrieval from multiple collections distributed across one or more processors. A number of efficiency bottlenecks develop, however, when the size of the databases is very large. Deciding which subcollections to search can address part of the problem, but there are other problems associated with the fundamental efficiency of the processes involved and the use of distributed resources. Image indexing and retrieval tends to make all of these problems worse since the databases and indexes are considerably larger.

#### General Methodology

The architectures and algorithms produced in this task will be evaluated using a combination of standard performance (efficiency) measures and effectiveness measures. The efficiency tests will be done using large PTO databases, including images, and the collection selection algorithms will be evaluated using the text subcollections of the patents.

## Technical Results

We have evaluated the multi-threaded version of INQUERY and found little performance difference compared to a client-server version. Experiments continue in this area and we have also been evaluating INQUERY's performance on large, distributed collections. This evaluation has shown that there are significant performance issues that need to be addressed in order to maintain adequate response times, although query optimiation and phrase-based indexing are expected to have a major impact.

## Important Findings and Conclusions

Initial results suggest no advantage to multi-threaded implementations in large, distributed environments.

## Significant Hardware Development

Purchases were made of disk to store large collections of patent data.

## Special Comments

Previous comments on fast network access to other PTO sites are particularly relevant here, since this will be required to both test the distributed architecture and to index and retrieve the full versions of the PTO databases.

## Implications for Further Research

We are still planning to ramp up the experiments on collection selection and result merging as well as continuing performance experiments.

**PTO Quarterly Report  
Distribution List  
w/Related Papers Attached  
1/14/97**

ARPA Agent: Harry Koch  
ESC/AXS w/papers  
Bldg 1704, Rm 114  
5 Eglin Street  
Hanscom AFB, MA 01831-2116

ARPA/ITO  
ATTN: Gary Koob w/papers  
3701 N Fairfax Drive  
Arlington, VA 22203-1714

ARPA/Technical Library  
3701 N Fairfax Drive  
Arlington, VA 22203-6145

Defense Technical Information Center (DTIC)  
Cameron Station  
Alexandria, VA 22034-6145

ESC/ENK  
ATTN: Ms Carole Stephan  
Bldg 1704, Rm 119  
5 Eglin Street  
Hanscom AFB, MA 01731-2116  
(Letter of Transmittal Only)

Lawrence J. Cogut  
Office of System Architecture & Engineering  
U.S. Department of Commerce w/papers  
Patent and Trademark Office  
Crystal Park 2, Suite 1004  
2121 Crystal Drive  
Arlington, VA 22202  
ph. (703) 305-8685  
fax (703) 305-9216  
lcogut@uspto.gov